VANDERBILT
Law School

Vanderbilt
Policy Accelerator
for Political Economy & Regulation

STATEMENT OF GANESH SITARAMAN

New York Alumni Chancellor Chair, Vanderbilt Law School
Director, Vanderbilt Policy Accelerator for Political Economy and Regulation

U.S. Senate AI Insight Forum, November 2023

Thank you for inviting me to join the AI Insight Forum. I am grateful for the opportunity to speak with you about the governance of artificial intelligence. As a law professor at Vanderbilt University, I study and teach constitutional law, foreign relations law, and economic policy and regulation – the regulatory process, the design of governance systems, and the regulation of networks, platforms, and utility (NPU) services. My research and teaching cover contemporary and historical systems of regulation, across a wide range of sectors, including tech platforms. I am also the director of the Vanderbilt Policy Accelerator for Political Economy and Regulation, which works on these issues.[1] This statement represents my views and not those of Vanderbilt University.

THE AI TECH STACK

In the public debates over AI, there has been comparatively little discussion of AI's industrial organization and market structure. This is surprising because critical layers in the AI technology stack are already highly concentrated. As in other areas, monopoly and oligopoly in these industries can not only distort markets, chill investment, and hamper innovation, but also facilitate downstream harms to users, threaten national security and resilience, and help accumulate private wealth and power in relatively few hands.

Understanding the AI technology stack offers a helpful framework for identifying what problems exist and where, and thus, what kind of regulation is needed and where. AI's technology stack can be described in four basic layers: hardware, cloud computing, models, and applications.

- Hardware. The hardware layer includes the production of microchips and processors—the horsepower behind AI's computations. This layer is extremely concentrated, with a few firms dominating important aspects of production.
- Cloud Computing. The cloud computing layer consists of the computational infrastructure that is required to host the data, models, and applications that comprise AI's algorithmic outputs. This layer, too, is highly concentrated, with three firms dominating the market. It features several dynamics that tend toward concentration and

[1] The remainder of this statement draws heavily upon Ganesh Sitaraman & Tejas Narechania, *Antimonopoly Tools for Regulating Artificial Intelligence*, Vand. Pol. Accelerator, Oct. 2023. For a longer treatment of these same themes, see Ganesh Sitaraman & Tejas Narechania, *An Antimonopoly Approach to Governing Artificial Intelligence*, Vand. Pol. Accelerator (Oct. 2023).

make sustaining competition difficult, including extremely high capital costs and significant switching costs to move from one provider to another.

- Models. The model layer includes data, stored in unstructured "data lakes" or more structured "data warehouses"; algorithmic models, which many think of as "AI"; and modes of accessing these models, including model hubs (where developers can download and use publicly available models) and application programming interfaces (or APIs, which allow developers to communicate with proprietary models that may not be publicly available). Some firms are fully integrated and offer all three products, which are then used to develop proprietary applications; others only offer models; and still others are more disaggregated, offering raw data or serving only as a model hub.
- Applications. Applications are the part of the sector that the public interacts with most directly. While some firms in the application layer build their products on open-source models, others offer applications built upon existing proprietary foundation models. Some firms are vertically-integrated, offering both the foundational model and applications built on them. Critically, though both types of firms compete in the applications market, those that are not vertically-integrated depend upon the firms that are for access to their models.

POTENTIAL DANGERS

Understanding the AI technology stack shows that significant portions of AI's industrial organization and market structure are likely to be, and already are, dominated by a small number of firms—and that these dominant firms are vertically-integrating across the stack. This concentration—an AI oligopoly—is concerning for a variety of reasons, including abuses of power, national security and resilience risks, exacerbated economic inequality, and its effects on democracy.

Abuses of Power. Concentration across the AI stack creates opportunities and incentives for dominant firms to abuse their power, with consequences for competitors, would-be entrants, and the public. These abuses could include, but are not limited to:

- High Prices. In hardware, dominant firms could demand monopoly and/or oligopoly prices for photolithography equipment, chip design, and chip manufacturing. Cloud computing firms might charge monopoly or oligopoly prices. In the model layer, the high costs of obtaining good data and sufficient compute infrastructure constitute a steep barrier to entry, and foundation model providers might therefore be able to raise prices to downstream application developers for model access.
- Self-Preferencing and Discriminatory Prices and Terms. Monopoly or oligopoly firms at each layer in the stack may discriminate between downstream firms, offering better terms or prices to their vertically-integrated businesses as opposed to competitors. This eliminates a level playing field and poses substantial downstream risks to competition and innovation. Cloud or model providers, for example, could favor their own applications over others by charging higher rates to third-party developers, exclude some third-party applications altogether, or find other ways to disadvantage third-parties.
- Lock-In Effects. Cloud services already feature lock-in effects that raise the costs for consumers to switch providers through egress fees and multi-year contracts. These effects

exacerbate the already-high switching costs in compute, due to factors like personnel expertise in a particular platform.

- Copying. Firms that copy applications from competitors and incorporate them into their own offerings prevent competition and chill innovation. Venture capitalists describe this practice as creating a "kill zone," wherein the likelihood of copying or acquisition by a dominant firm discourages investment in innovative companies and products.

National Security and Resilience. With very few chip companies, the possibility that one foundry or cloud provider could be shut down due to a war, pandemic, weather event, or other emergency is significant. Just as concerning is that faulty foundation models, if offered by only one firm, can lead to widespread error that could be catastrophic in emergency situations. More broadly, foreign ownership—even if partial—over critical technology resources raises national security questions.

Economic Inequality. Like concentrated power in other industries, an AI oligopoly is likely to further economic inequality. Concentration means that a small number of firms will capture the vast majority of the sector's profits. And while it is too early to tell exactly how the introduction of AI at scale will change labor markets, it could very well lead to job losses and widening inequality, especially in the medium-term.

Constitutionalism. An AI oligopoly can also contribute to the erosion of our constitutional system. Concentration in AI may give a relatively small number of companies an outsized influence over the information ecosystem. Economic power also often translates into political power, and may leave the federal government dependent on a small number of firms and individuals for critical resources.

POTENTIAL SOLUTIONS

The lower levels of the AI stack, especially cloud and models, are like electricity: they are essential utilities that power a wide range of activities. They also feature high capital costs and economies of scale. Indeed, machine learning itself has the characteristics of a natural monopoly, even under narrow definitions.[2] In other similar sectors—transportation, telecommunications, energy, banking—the longstanding American way was to regulate firms with these characteristics using utility-like tools.[3] Congress passed comprehensive statutes that addressed a wide range of challenges in each sector—including the downsides of oligopoly and national security risks—and tailored the tools to the features of that sector. Congress should again consider taking this approach, including considering the following tools that have applied in other sectors.

Structural Separations. Structural separations "limit the lines of business in which a firm can engage."[4] The central benefit of structural separations is that they prevent a business from self-preferencing or leveraging their power from one business-line into another. In addition to

[2] Tejas Narechania, *Machine Learning as a Natural Monopoly*, 107 IOWA L. REV. 1543 (2022).
[3] For a discussion, see, e.g., MORGAN RICKS, GANESH SITARAMAN, SHELLEY WELTON & LEV MENAND, NETWORKS, PLATFORMS, AND UTILITIES: LAW AND POLICY (2022).
[4] ID. at 28.

preventing conflicts of interest and leveraging profits to gain competitive advantage in another line of business, structural separations also limit the concentration of economic power and promote a diverse business ecosystem of users of the platform. With respect to AI, structurally separating the cloud layer from higher layers in the stack could address a wide range of market dominance problems identified above. It would treat cloud computing platforms as utility providers of a service (namely, computational capacity) that is open for all kinds of uses and ensure that those providers cannot prioritize their own downstream business lines over their competitors'.

Nondiscrimination, Open Access, and Rate Requirements. A complement to structural separation requirements are nondiscrimination and equal access rules, sometimes coupled with rate regulation.[5] Nondiscrimination rules require a firm to treat downstream businesses neutrally, including its own vertically-integrated business lines if there is no structural separation in place. Equal pricing rules are an essential corollary to nondiscriminatory rules because firms could charge prohibitive prices as a workaround to evade their open access obligation. In some areas, regulators have also directly set the prices firms can charge. Rate setting "is usually directed toward preventing NPU enterprises from lowering output and raising prices," while simultaneously ensuring firms earn a reasonable return on invested capital.[6] At the cloud level, cloud providers should be required to treat all downstream businesses in a nondiscriminatory fashion, be open to all comers, and offer transparent, uniform, publicly-available prices. Foundation models and APIs could also be subject to such rules, so that app developers have reliable access to these resources.

Interoperability Rules. Interoperability rules lower barriers to entry and thus stimulate competition by "allowing new competitors to share in existing investments" and "imposing sharing requirements on market participants."[7] In the telecommunications context, for example, rules that required a telephone provider to transfer a user's phone number to a competing provider (and thus required that the providers work together on an interoperable number portability system) facilitate competition by reducing switching costs for users. Those rules targeted a notable lock-in effect: It is quite cumbersome to let all your contacts know you have a new phone number. In the AI stack, interoperability among cloud platforms could be improved, easing transitions from one provider's system to another.

Entry Restrictions and Licensing Requirements. These rules limit entry into a sector to firms that have registered with an appropriate regulator or otherwise have approval from the government (often in the form of a license or certificate), and are usually justified on one (or more) of three different grounds:[8] (1) Entry restrictions can ensure safety and reliability. By placing conditions on entry into a sector, regulators can ensure that firms will operate safely and effectively. (2) In some markets (particularly those typically characterized as natural monopolies or oligopolies) competition can lead to waste and ultimately deter capital investment. Entry restriction can prevent these downsides, creating a stable environment for capital investment. (3) In sectors in which universal service—i.e., ensuring that everyone can access the regulated

[5] ID. at 24-26.
[6] ID. at 25.
[7] Narechania, *supra* note 2, at 1555.
[8] RICKS ET AL., *supra* note 3, at 29-30.

service—is a critical policy goal, regulators will often limit entry to the market. This is because open competition often undermines universal service policy goals. In the AI stack, entry restrictions might be deployed to ensure that certain foundation models and their associated applications are effective, and do not pose substantial risks to health and safety, or of bias. Similarly, licensing rules could oblige cloud and model providers to "know their customers," as in banking law, and ensure that entities in the model layer have checks in place to ensure non-discriminatory access, fair pricing, and safety. Applications could be required to register with the model or cloud they use, to make it easier to identify and address dangerous or problematic behavior on a post hoc basis. Entry restrictions could also prevent foreign ownership of firms at different layers in the AI tech stack. Policymakers should avoid getting caught into all-or-nothing thinking, as there are many ways to design such requirements and tradeoffs in those designs.

Public Sector Capacity. Building up public sector capacity can also ensure competition, increase access to scarce resources for researchers and scientists and government uses of AI that may not be profitable to private firms, and ensure national security needs.

- Publicly-Run Cloud Infrastructure: A publicly-run cloud service would allow government access to this essential infrastructure, without paying oligopoly prices, getting locked into a single system, or further entrenching dominant players with added funding and access.
- Public Data Resources: Data is also foundational for AI applications, and it is a resource that depends on extraordinary scale. If leading data sources are all proprietary, then the companies that control them could raise prices on downstream businesses or researchers who rely on that data for their models or applications or even deny them access, perhaps if they seek to develop a competitive product. Public data resources could alleviate these problems.
- Federal Government AI Personnel. The federal government could hire many more AI specialists to help integrate AI into federal programs, in the process improving service delivery and efficiency across the government. These will need to be highly-talented, skilled individuals who are also in demand in the private sector. Building a new and notable system—like a U.S. AI Service or U.S. Technology Administration—to supercharge federal AI capacity would ensure government can take advantage of the benefits of AI, and also not be so dependent on contractors, consultants, and the biggest technology companies.[9]

CONCLUSION

Understanding the AI tech stack reveals that there are dangers at each layer and across layers. Public policy can address these dangers, and lawmakers should consider using the tools at their disposal to do so. These tools should be adopted carefully and ideally be designed in an integrated fashion. This would lead to a system whose parts work together and which ensures that artificial intelligence develops and is deployed in beneficial ways.

[9] For more on this idea, see Ganesh Sitaraman & Ramsay Eyre, *Building Public Capacity on Artificial Intelligence*, Vand. Pol. Accelerator, Oct. 2023.